My work and research experience spread across a number of areas in computer science, mathematics, and biology. My strengths lie in taking practical, application-motivated problems, abstracting them to discrete computational models, and solving them using fundamental techniques in algorithm design and combinatorics. I am particularly interested in working on biological problems.

My Ph.D. research focused on problems in multichannel communication. I modeled them as classical graph-theoretic problems and provided the best solution to date to a 30-year open problem in graph theory and devised a new model and technique for multichannel communication encoding design.

Slightly over a year of my postdoctoral fellowship was focused on problems in computational biology, specifically phylogeny reconstruction and design of mating strategies for controlled animal breeding programs. I also have collaborated with the Sandia National Laboratory on the deployment of air and water quality sensors to ensure contamination detection.

Phylogeny Reconstruction - Postdoc

My work in phylogeny reconstruction is a mixture of both theoretical and experimental research. It focuses on various aspects of combining phylogenetic information from multiple sources. This work is being conducted in collaboration with the research groups of Bernard Moret and of Tandy Warnow, and with biology collaborators at University of Texas at Austin and elsewhere.

One of the fundamental problems in biology is reconstructing the evolutionary history of a set of organisms (or taxa). Relationships among the taxa are modeled as a phylogenetic tree. Generally, multiple trees exist for any set of taxa. One source of this inconsistency is the fact that computational phylogeny reconstruction heuristics return numerous near-optimal trees with the same optimality score. Multiple phylogenetic hypotheses are also a result of different types of data being used or different taxa representatives being chosen for an analysis. Having a set of trees as an answer is quite unsatisfying since the objective is to obtain a single "representative" tree that best describes the relationships among the taxa of interest. Commonly, a *consensus* or a *supertree method* is applied to combine all of the trees into a final tree.

Are Near-Optimal Trees Good Enough? Since the outcome of a phylogeny reconstruction process is a consensus of the best-scoring trees, is there a significant difference between optimal and near-optimal outcomes? Can we save time by stopping the tree search at sub-optimal scores? Is there a significant quality difference between results with different types of consensus? We are currently performing large-scale experiments to answer these questions. The preliminary results indicate that if majority consensus is used to combine top-scoring trees to return a single answer, it is indeed possible to stop the phylogeny heuristic search at an earlier stage than currently done, without sacrificing the quality of the answer. This would result in significant time savings, since in current large-scale phylogenetic analyses it may take hours to improve the score by just one. To develop reliable stopping criteria, we need to be able to calculate many of the statistics and the consensus of the current best-scoring trees on the fly. In [6]* we presented optimal on-line algorithms for calculating the two most common consensus methods, strict and majority-rule. We are working on the algorithms for the rest of the necessary statistics.

Comparison of Consensus Methods. The practical and quantitative behavior of the consensus methods is poorly characterized. We have performed the first large-scale analysis of the three most commonly used consensus methods—strict, majority-rule, and greedy consensus—in an experimental environment [4]. We have shown that, surprisingly, the greedy consensus, while commonly used to produce binary trees from majority-rule consensus, sacrifices accuracy to do so. We have also provided a theoretical justification for this result.

*Citation numbers refer to the publication list in curriculum vitæ

Population Biology - Postdoc

Modeling is commonly used in population biology to answer various biological questions. Presently, most of the modeling, including computer modeling, is continuous and stochastic. However, in many cases the empirical statistical information underlying a continuous model is very hard to obtain: it is often collected for a different animal population, the population is too small and has large variance, or the environmental conditions have changed since the information was last collected. For these and many other biological and mathematical reasons, continuous numerical modeling is not robust or simply impossible.

In collaboration with Cristopher Moore, Alexander Russell, and Jared Saia, I used the problem of designing and comparing animal breeding strategies as a test for discrete modeling techniques. The aim of a breeding program is, beginning with a small population of known individuals, to create or maintain certain genetic characteristics within the population. Heuristic mating strategies are the norm in practice. Although some analysis and comparison of these heuristics is done using stochastic modeling, the models are not well-suited for answering the main qualitative question, "which strategy is better?"– inherently a question of algorithm analysis. In [5] we proposed the first discrete computational model of the controlled breeding problem and analyzed common mating heuristics for two specific objectives. This simple discrete model of the breeding problem provides a novel, viable and robust approach to designing and comparing breeding strategies in captive populations.

Air Quality Sensor Placement

Monitoring air and water supply is a significant safety measure. Where should sensors be placed so that it is always possible to detect contamination and locate the contamination source? Previously, this family of problems has been approached with a continuous flow model using numerical data which is often noisy and hard to obtain. In collaboration with William Hart and Jared Saia, I proposed the first discrete, graph-based model for the sensor placement problem [7]. This model is more robust and potentially allows faster solutions than the continuous approach. We analyzed the complexity of several variations of the problem and developed exact and approximation algorithms.

Multichannel Communication - Ph.D.

One way of dealing with the rising volume of communication over media with limited bandwidth is to break up the data into pieces. This approach is currently used in various settings such as packet switched networks, various multimedia protocols, and wireless and mobile computing. In the particular setting I have investigated the communication medium is considered unreliable, and in case of failure the information sent over it may be corrupted, lost, or delayed and rendered useless. The goal is to design an encoding scheme that breaks up the information in a way that, in case of communication failure, minimizes the error after reassembly. This question was posed in an information-theoretical context as the Multiple Description problem in 1979. It is a generalization of Shannon's classical problem of source coding subject to a fidelity criterion. Previously, there have been no optimal constructive results, indeed, no constructive results for encoding into more than two pieces. I was able to show that, depending on the type of data, error measure, and type of failure, the problem of designing these encoding schemes is equivalent to several classical problems in graph theory. These problems are vertex-labeling problems in product graphs. Many of these have been open for over 30 years. In [1], [2], and [8] we demonstrated lower bounds and provided nearly optimal or optimal encodings of data into an arbitrary number of pieces for two specific configurations of the problem and proposed the best encoding for yet another. The techniques I have developed are already being used successfully to solve other problems in multichannel communication (for example, in works of Servetto, Balogh, and Csirik).

Future Work

I plan to continue my research in computational biology, building on what I consider to be my strength of developing clean discrete computational models for real-world problems. The two main areas of research I plan to address are (i) developing methods for phylogeny reconstruction that use heterogeneous data and (ii) developing discrete models and techniques in population biology and ecology.

Heterogeneous data in phylogeny reconstruction. My main research agenda in phylogeny reconstruction is to develop phylogenetic reconstruction methods that can handle heterogeneous data. All computational methods today intrinsically rely on a measure of difference between two species, a concept of evolutionary distance, defined on a single type of data. To reconstruct the entire Tree of Life, we must use all available data, including DNA, genomes, paleontology, morphology, geography, and any recorded evolutionary history.

Much of the existing non-molecular data can be formulated in terms of constraints that can be positive (group species together), negative (forbid certain groupings), temporal (define a partial order), etc. A method capable of reconstructing phylogenies based on such constraints could easily combine heterogeneous data.

Consensus and supertree methods can be viewed as examples of constraint-based reconstruction methods that use positive constraints. In collaboration with Tandy Warnow, I am working on classification of the existing consensus methods and their extensions according to the inference rules for the various types of constraints. A logical continuation of this work lies in developing new algorithms for combining these various constraints, since no current methods to do this exist. I also hope to generalize these methods to handle phylogenetic networks, as well as trees, since temporal constraints and conflicting constraints are crucial in network reconstruction.

Near-optimal trees in phylogenetic analysis. Our current experiments on biological data show that, if a consensus of the best trees is returned as a final answer, then there may be no significant difference between the highest scored trees and the answer obtained from the trees with lower scores. This can have a significant impact on the running time of the phylogeny reconstruction heuristics, but it is necessary to devise and test criteria for the termination of the heuristic search. Some possible such criteria include: lack of change in the majority consensus tree of the current best scored trees over a period of time and lack of change in the maximum and average pairwise distance amongst the best-scored trees. Our on-line consensus algorithms [6] are the first step in this direction.

Discrete population biology. Our work on controlled animal breeding strategies clearly demonstrates that discrete approaches can be beneficial in population biology. Many other biological problems can be formulated as optimization problems and solved using discrete computational techniques: What is the smallest number of animals from a closely related population needed to stabilize a given population? What is the maximum number of animals from a closely related population that may be introduced so that the genetic uniqueness of a given population is preserved? Will a population survive the introduction of a physical barrier? If there is a level of interaction between two populations, at what point do they become effectively one population? I plan to work with biologists on problems of this flavor, using discrete techniques.

Epidemiology. Population biology and phylogeny are closely related to epidemiology. When a disease is not easily transmitted, standard diffusion models do not work very well to identify populations at risk. Continuous models are also not very good at identifying possible routes of transmission. I plan to collaborate with epidemiologists to develop discrete epidemiological models for such cases.